20 years of UK Budget speeches: correspondence analysis vs. networks of n-grams

Andrea Gobbo[1], Fahime Same[2]

[1]LSE, London – andrea.gobbo.lse@gmail.com

[2]University of Cologne – f.same@uni-koeln.de

Abstract

In computational linguistics, repeated segments in a text can be derived and named in various ways. The notion of n-gram is very popular but retains a mathematical flavor that is ill-suited to hermeneutical efforts.
On the other hand, corpus linguistics uses many definitions for multi-words expressions, each one suited to a specific field of linguistics.
The present paper contributes to the debate on term extraction, a particular notion regarding multiword-expressions that is widely used in summarization of scientific texts. We compare and contrast two methods of term extraction: distance-based maps vs. graph-based maps. The use of network theory for the automated analysis of texts is here expanded to include the concept of community around newly identified keywords.

Résumé

Dans la linguistique computationnelle, on peut dériver et nommer les segments répétés de texte de plusieurs façons. La notion de n-gram est aujourd‟hui assez populaire, mais elle retient une nuance mathematique qui mal s‟adapte aux efforts herméneutiques. D‟autre part la linguistique des corpus emploie plusieurs définitions pour les expressions multi-mots, chacune s‟adaptant a un domaine spécifique de la linguistique. Cette communication contribue au débat sur l‟extraction des termes, une notion spécifique qui est fort employée dans la récapitulation des textes scientifiques. Ici on compare deux méthodes pour l‟extraction des termes: les cartographies basées sur la distance statistique et les schémas basées sur les réseaux (ou graphes). L'emploi de la theorie des réseaux pour l‟analyse automatisée des textes est ici étendue jusqu‟à inclure les mots clés dans les communautés.

Keywords : budget speech, co-occurrence, collocation, n-gram, keyword, text networks.

1. Introduction

The work presented here is a methodological exercise on an ad-hoc constructed corpus of 173,605 occurrences. The aim of the textual analysis is to extract collocations with statistically significant co-occurrence; also they must be used plausibly by competent speakers of English language. The theoretical section 2.2 will explain why these two dimensions, significance and plausibility, have to converge in order to validate the method through semantic interpretation.

We propose two distinct analyses of the same corpus and of the single speeches. Firstly, we perform a correspondence analysis of the combined speeches in order to show how a classic multidimensional diagram being extracted through the Iramuteq software is able to visualize semantic classes. The second analysis is the application of network analysis to patterns of token co-occurrence derived from a parser program (Vos-viewer). By comparing and contrasting the two approaches, we aim at elucidating how factor and correspondence analysis methods (Distance-based maps) lead to a different exploratory power compared with text

network analysis (Graph-based maps) which are nonetheless based on the same contingency matrices (Jan van Eck et al., 2008).

In particular, it will be shown how the second method of network analysis gives new insights into the interpretation of keywords emerging from the analysis of a subset of n-grams and of collocations.

2. Theory: distinguishing n-grams, collocations and terms (keywords)

The need to distinguish between different concepts arises from the fact that in different disciplines such as information extraction, text mining and corpus linguistics, they carry subtly different meanings. The case of n-grams is particularly paradigmatic. The notion of n-gram was first introduced in theory of stochastic processes and communication theory (Markov A.A. & Liebmann H., 1912; Shannon, C. E., 1951), but later found its way into computational linguistics permeating recent debates on keyword recognition. On the other hand, the idea of *keyword* is nowadays massively found on the internet. The keywords we will end up with will characterize knowledge circles of insiders who deal with specific topics (Leydesdorff, 2009).

From the computational point of view, keywords are n-grams. However, they also carry important information from the linguistic point of view. As a result, it seems necessary to reconcile computational and linguistic insights, hence to come up with a definition which is acceptable and applicable in both domains.

When dealing with natural language processing, "n-gram" is a very big set of tokens: significant collocations in everyday language are just a small subset of those n-grams. It might be useful to imagine a cascade combination of such words associations. The meaningfulness of subsets is humanly gauged against public discourse; in this regard the larger set contains all statistically significant tokens whereas the smaller set contains only the collocations that are most relevant for a certain public in a certain time period. In the corpus linguistics literature, there are plenty of definitions for significant collocations. For example, Firth (1957) introduced the notion of linguistic gestalt, later used by Lakoff (1976). Other names given to significant collocations are *formulaic expressions* or *phrases*, *polywords*, *ready-made units* among many other less usual ones. They can be described as crystallizations of ways of saying something that enter the literature and the public talk from many different channels. A general definition of this concept is the following: "a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray, 2002:9).

It can be further argued that patterns of significant usage frequencies can refer back to socio-linguistic phenomena like the common use of a collocation in social groups (synchronic corpora) or even it could be used to reconstruct patterns of diffusion of innovations in circles of insiders across time (diachronic corpora).

2.1. Keywords as special cases of collocations

One of the fundamental steps for processing natural language is terminological or keyword extraction. With the growing availability of online texts and corpora, terminological extraction is regarded as a highly-potential method for domain-specific information retrieval, also for creating domain ontologies. Identifying N-grams as a consecutive occurrence of N-tokens in a string are an essential procedure for textual information extraction. Their

importance in domain-related research is due to the fact that collocations could be derived from n-grams. To put it simply, collocations are a subset of n-grams.

From a linguistic point of view, the term collocation does not have a unified definition. "a collocation is any holistic lexical, lexico-grammatical or semantic unit normally composed of two or more words which exhibits minimal recurrence within a particular discourse community" (Siepmann, 2005). The important point about the collocations is that the co-occurrence of constituents in a „short span of text" is more often than chance **(**Seretan, 2011). Collocations" classification methods are also quite diverse. The proposed classifications are mainly based on three different criteria. The first classification concerns the syntactic installation of the constituents. The second one is based on the semantic features, i.e. seeing whether the constituents are in their literal or figurative sense; and the third one is about the accordance to the commutability of the collocation"s elements which means whether the elements can be replaced freely or they are restricted.

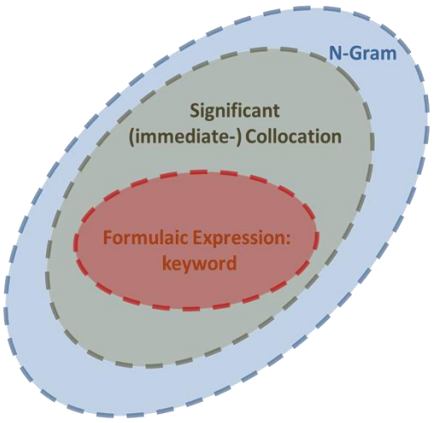

Figure 1. Conceptual nesting of multi-word expressions.

In the current work, we mostly focus on the first type of classification; i.e. the structural categories in a domain-specific field. Hausmann (1989: 1010) classifies collocations into six different structural categories: adjective + noun (heavy smoker), (subject-)noun + verb (storm – rage), noun + noun (piece of advice), adverb + adjective (deeply dis-appointed), verb+adverb (severely criticize), and verb + (object) noun (stand a chance) (Nesselhauf, 2005, 22).

Considering that the extracted collocations might be general and thus contain less information about a specific domain, it is important to extract the keywords in a specific domain such as the one we chose here. Extracting the keywords could be beneficial from different perspectives. First of all, it sheds lights on the dominant tendencies and intricacies of a domain. As an example, special needs of a domain might have an impact on the structure and length of the frequent keywords. Secondly, keywords manifest how experts tend to encode and transfer the information in a domain; also to what extent the keywords could penetrate into the public audience"s lexicon. Thirdly, the keywords themselves are able to find their way into a specific domain after being coined by a public community. Finally, the impact of special events in the domain could be easily traced with the help of analyzing the usage frequency of related keywords. As keywords play a crucial role in gaining more knowledge on a specific domain, it seems necessary to choose an appropriate method for extracting them. Keyword extraction helps us eliciting the most related terms and information from a corpus. This process could be classified into three different categories: statistical, linguistic and mixed methods. *Statistical methods* are mainly concerned with the non-linguistic information of a text; e.g. term frequencies, inverse frequency and the position of a keyword in a text. *Linguistic method*s are centered on syntactic functions of keywords and their semantic criteria (Das et al, 2013). *A mixed model* encompasses linguistic method with some statistical measures. In the current work, a mixed model has been chosen for extracting the data.

3. Iramuteq cluster and specificity analysis

We first propose the exploratory cluster analysis with Iramuteq (Ratinaud et Marchand, 2012). The reason to perform it is to get a bird‟s eye view of the corpus and demonstrate the significant words in a clear way; since the second analysis is more tentative, Iramuteq acted as a sort of cross-validation solution. Further, among its functionalities, Iramuteq outputs a list of repeated segments based on Khi2 significance level for each class. This feature is precious for a comparison with the similar function in Vos-viewer.

3.1. Material: corpus

The corpus used in this analysis is ad-hoc constructed. It contains 21 budget speeches given by four British Chancellors of the exchequer in front of the parliament assembly between 1994 and 2014 at the beginning of each fiscal year. For the two distinct analysis, we used both the corpus on its whole (Iramuteq preliminary analysis) and the single budget speeches (Repeated segments with iramuteq and terms extraction with Vos-viewer). It seems that speeches during the years have a tendency to become shorter.

service period	year	Chancellor	Speech word lenghth
4 ys	1994	Clarke	11.509
	1995	Clarke	9.297
	1996	Clarke	9.585
	1997	Clarke	7.595
10 ys	1998	Brown	8.322
	1999	Brown	9.101
	2000	Brown	7.781
	2001	Brown	8.178
	2002	Brown	9.120
	2003	Brown	9.167
	2004	Brown	8.372
	2005	Brown	6.868
	2006	Brown	8.714
	2007	Brown	7.042
3 ys	2008	Darling	7.642
	2009	Darling	7.395
	2010	Darling	8.783
4 ys	2011	Osborne	8.607
	2012	Osborne	8.176
	2013	Osborne	7.492
	2014	Osborne	7.726

Table 1. Corpus composition; 21 UK budget speeches. The table outlines author, date and length of the speech.

The 173,605 tokens of the complete corpus were classified according to author and year: (variables are: *year_, *chanc_). We first consider the variable *chanc_ and the bidimensional CA plane output. A straightforward interpretation shows that Brown‟s rhetoric is quite different from the others‟, being placed on the left of the 1st dimension axis while the other three chancellors are almost in line on the right of the same dimension. Osborne and Clarke in turn oppose each other on the 2nd dimension, the Y axis. Darling seems to use a more balanced vocabulary compared to these two.

Directing attention to the variable *year_, we see that adjacent years follow a clear pattern and cluster together in groups of 3 to 5 years. A probable Guttmann effect is also present in the data, whereby on one end we have the oldest speeches (Clarke) and on the other the more recent ones (Osborne). Also it is clear that Brown‟s discourse have been divided into two phases, the former until 2002 and the latter from 2003 to 2007 (in appendix).

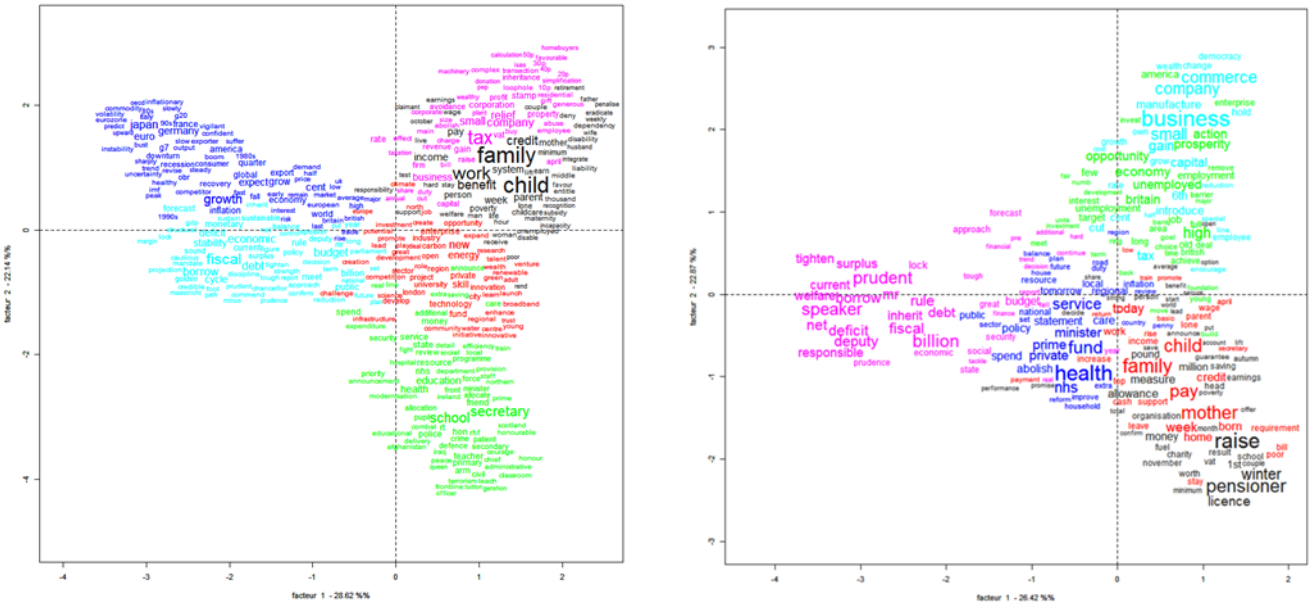

*Figure 2 and 3. FCA (factor correspondence analysis) based on the variables *chancellor and *year, obtained with GNEPA clustering algorithm: former on whole corpus, latter on the 2000 Brown speech.*

3.2. AFC analysis

Performing a descending clustering classification of text segments (GNEPA method) with default parameters (12 tokens text segments vs. 14 tokens, with a maximum of 10 classes), we obtain 6 semantic classes. This kind of analysis often generates classes that are difficult to read due to the amount of words included. Even by reducing the number of forms, we hardly get a better picture, because the algorithm does not take into account the syntactic position of the constituents inside the text, so that clusters are not easily interpretable.

In our opinion, the challenging issue is the identification of collocations and keywords; by reducing the number of forms during the analysis, we decrease the possibility of getting auxiliary and function forms inside keywords. As an example, let us consider two GNEPA analysis on the whole corpus and on a single speech. By increasing the size of the corpus, as the number of tokens increases, the image gets more blurred and the identification of related keywords becomes more difficult. On the other hand, by zooming into a single corpus, we would be able to identify the topics of interest in a specific period of time; at this level, we begin to notice what words might co-occur at the same sentence and build plausible collocations.

3.3 Repeated segments

Following our reasoning about different sets of n-grams, collocations down to terms and keywords. Let us consider, as an example, the third class extracted in Brown‟s speech. Iramuteq can give as its output repeated segments ordered by class specificity. At this stage, the program handles such repeated segments as pure n-grams; i.e. it does not operate POS tagging. If the syntactic position is not taken into account, the observer would have difficulties in seeing which ones could be plausible keywords. The program‟s functionality brings us closer to the concept of term extraction. The fourth class, for example, contains

2019 repeated segments (from bi-grams to 10-grams). To scale down the amount of information, we can look at the tokens with P-value of <0,0001:. In this case, we will have 419 tokens; however this number is still too big for a quick bird"s eye view of topics. As we notice in the list of significant forms below, some collocations are more interesting than others; their form is more concise and they convey more information. As an example, the collocation "economic-cycle" encompasses more information compared with "the economic" or "the economic cycle". In line with recent literature we propose to call the first one a *keyword*, whereas the others would be *naïve collocations* or *usage collocation.* Of course the distinction we propose is a matter of heuristic interpretation and will be clarified in the next section.

CHD | Profils | AFC | **Profils des segements répétés** ×

classe 1 | classe 2 | classe 3 | **classe 4** | classe 5 | classe 6

num	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	44	49	89.8	155.96		fiscal rules	< 0,0001
1	41	45	91.11	148.62		our fiscal	< 0,0001
2	40	45	88.89	139.56		economic cycle	< 0,0001
3	36	39	92.31	133.12		monetary policy	< 0,0001
4	33	35	94.29	126.03		the economic cycle	< 0,0001
5	30	33	90.91	108.36		fiscal policy	< 0,0001
6	37	49	75.51	99.12		public finances	< 0,0001
7	23	23	100.0	95.92		current budget	< 0,0001
8	28	32	87.5	95.3		our fiscal rules	< 0,0001
9	49	78	62.82	94.57		national income	< 0,0001
10	40	58	68.97	91.59		the economic	< 0,0001
11	39	56	69.64	90.86		of national income	< 0,0001
12	23	24	95.83	90.02		inflation target	< 0,0001
13	41	61	67.21	89.64		of national	< 0,0001
14	21	21	100.0	87.58		golden rule	< 0,0001

Table 2. Repeated segments belonging to the fourth class.

4. Networks of n-grams

We have by now realized that not all parts of speech are used for building specialist word combinations; as a consequence, keywords show a preferential grammatical structure. It is well known that most keywords in scientific papers are made up of adjective+noun or noun+noun in different lengths. Recent one but also some older literature refers to these special N-grams as "terms" (Leydesdorff 2009). Good candidates to form keywords are the collocations or N-grams composed as: Adj+Noun; Adj+Adj+Noun; Adj+Noun+Noun; Noun+Adj+Noun… (bi-grams; tri-grams; 4-grams, etc.).

4.1. Term extraction with Vos-viewer

The program VOS (visualization-of-similarities) is composed of three parts: 1) a POS tagger and a selection filter, 2) a proximity detection algorithm based on co-occurrences, 3) a tool for the calculation of similarities based on Euclidean distances as an alternative to multidimensional scaling. We used the first two features of the program in order to generate networks of keywords. Ultimately, we use the program as a parser for selecting the terms

which have a significant co-occurrence in a text. This latter procedure is similar to a TFIDF evaluation.

Networks of single words and combined keywords are computed as a relation between co-occurrence of each item in the whole text and as an adjacency relation to another term. The association measure is the *Pointwise Mutual Information (PMI)*. This similarity measure normalizes the adjacency matrix of data delivering a proximity index. As is detailed in the program"s handbook, significant keywords are assumed to be more specific (thus significant) if they have a higher PMI score (Van Eck, N. J., & Waltman, L., 2011)

4.2. Networks of terms

Terms, either as a single word or as an n-gram, are eventually computed into a .paj file. The first part of the file contains a matrix that attributes a progressive number to each item. The latter part keeps track of how items combine, involving also the number of edges to each combination. The network is undirected, that is, it does not take into account if words precede or succeed each other.

In this paper we include some depictions of the networks of terms obtained. We consider three different outputs of the Brown-2000 speech. The first is a Vos-viewer network that automatically chooses the most significant terms to show.

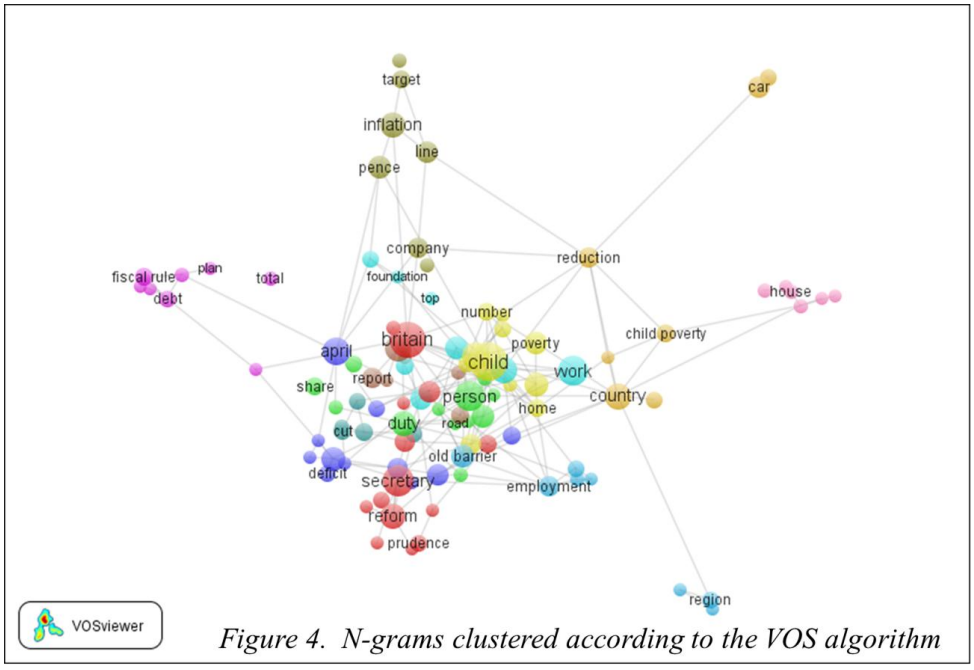

Figure 4. N-grams clustered according to the VOS algorithm

In our opinion, the reading of clusters in this depiction is not optimal; for this reason, we also used the .paj file of the same network as an input to Pajek, the well-known network visualization program from the University of Ljubljana. The route-mappers algorithms of Pajek allow extracting communities from such files. We obtained the depiction below from which we further took out the links between communities. What we were left with was a constellation of *keywords communities* that very well identify topics in the speech.

Figure 6 depicts the final stage of analysis. The rather long stepwise analysis we devised brought the novel result of getting networks of terms out of a clustering procedure. By

handling networks instead of clusters (graph-based maps instead of distance-based ones), we can derive some neat communities that in our opinion clarify what discourse topics could be.

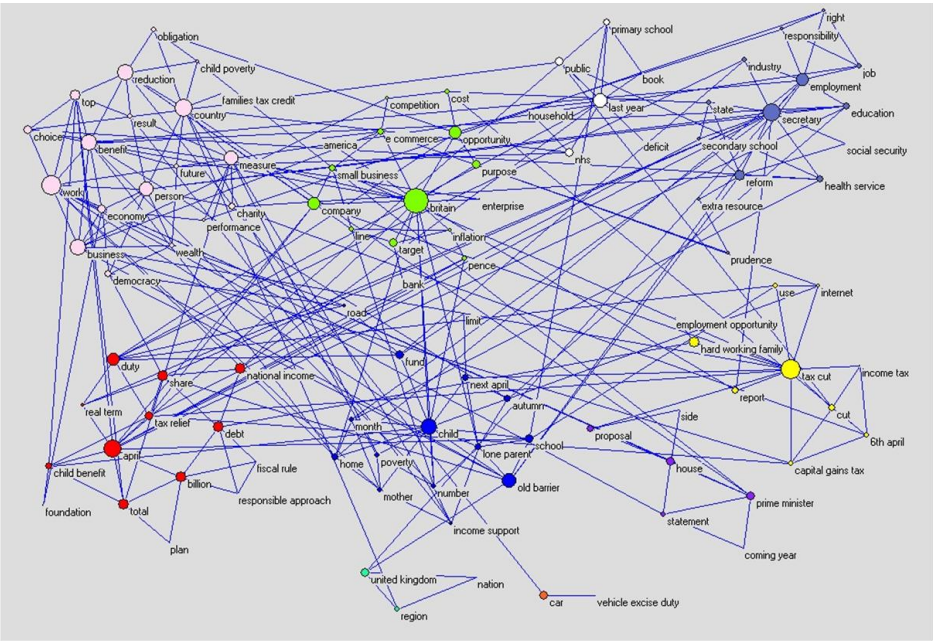

Figure 5. Budget_2000_Brown-visualized with Pajek: whole network: Communities with betweenness centrality

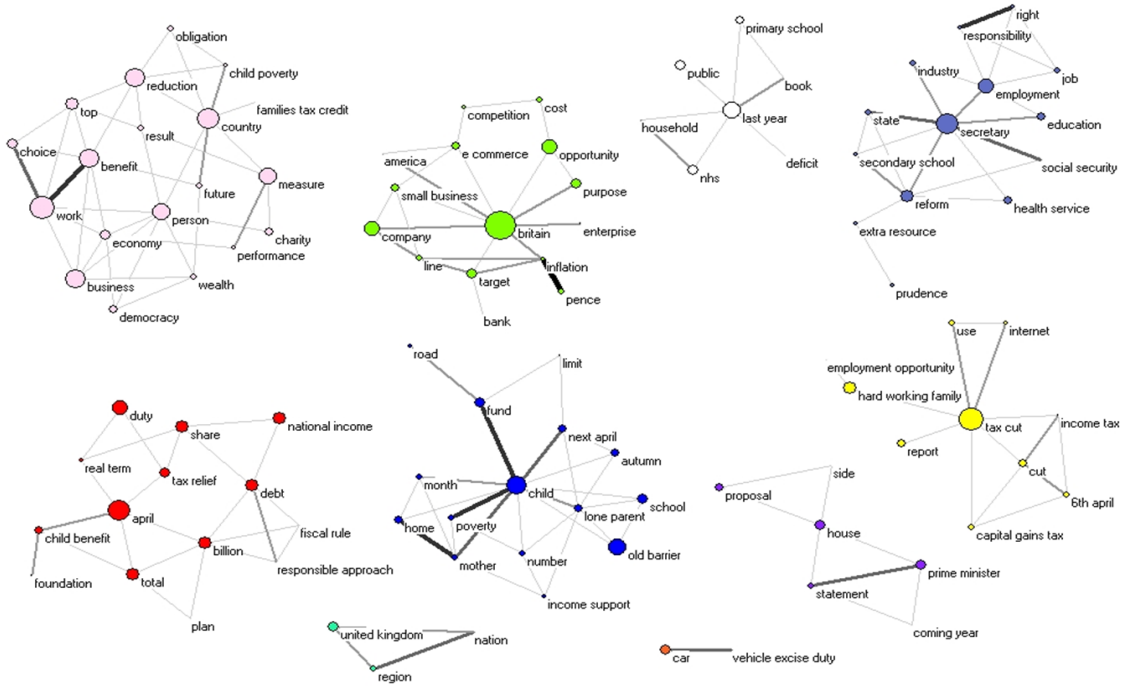

Figure 6. Budget_2000_Brown-visualized with Pajek: communities without external edges and betweenness centrality

5. Discussion: exploring the unexpected keywords

Drawing from network theory, some considerations can be advanced in relation to the use of networks for textual data. Significant collocations (keywords) as vertices in communities generally show small degrees (low connectivity). In a domain-specific text, keywords seem to follow the same pattern, they usually appear in one cluster and they are seldom central. This observation means that they appear isolated from other more popular words and that they need some supporting function from sets of frequent words. In turn, they provide us with a better view of topic in a text in a certain time. By that, we mean significant collocations might show a strong pattern of adoption-decay particularly explicit if placed inside organized sets of single words (Rogers, E. M., 1962).

5.1. Checking for trends with the Google books corpus

As a second-step validation, we tried to input some of the constructs both into Google-n-gram viewer and Google Trends. Whereas Google n-gram viewer might be more frequent in corpus linguistics studies, our insights were especially corroborated by Google Trends. In the analysis of the 2012 speech by Osborne, for example, a 3-gram keyword is interesting: *higher rate taxpayer*, strangely related with *family* and *child benefit*.

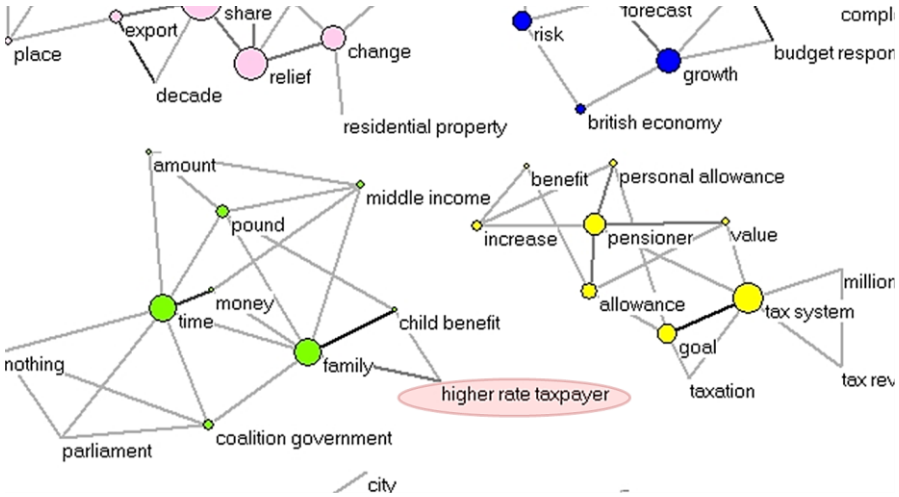

Figure 7. Budget_2012_Osborne, extract of communities

Google trends is dedicated to web searches of keywords usage. It embeds a data corpus of static pages from 2004 onwards. The added feature compared to n-grams viewer is the tracking of IPs, so that it is possible to a certain extent to know where the term was searched for. Needless to say, many of the keywords used in budget speeches had their epicenter in London. The keyword *higher rate taxpayer*, for example, showed a very interesting adoption curve: it appeared in August 2007, remained latent until October 2010, went into a short decay until exactly the budget speech month, when it regained its popularity and is constantly used thereafter.

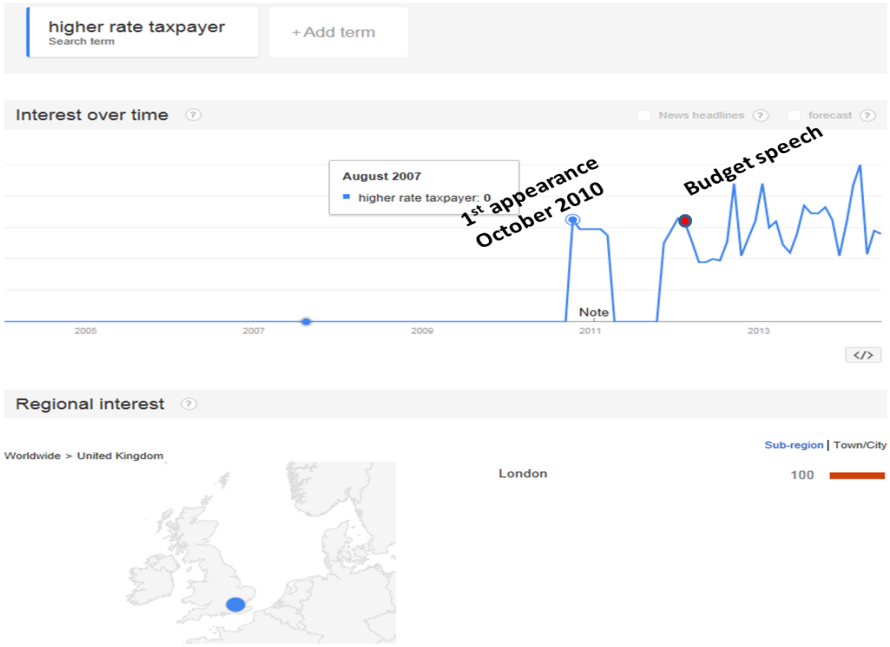

Figure 8. Time and location of a keyword appearance

5.2. Long term trends

Google Books N-gram Viewer contains more than 4 million printed texts from 1800 onwards. Checking for the frequency of n-grams in that repository equates to evaluate the long-term usage of a certain word in the most complete corpus of English language available (Michel, J.B. et al., 2011).

We checked as an example a keyword in one of Clarke"s budget speeches, enterprise investment scheme. This case is paradigmatic of an adoption dynamic; the budget speech once again fulfills the role of boosting the adoption of a very specific keyword. The function profile closely tracks the innovation-adoption sigmoid.

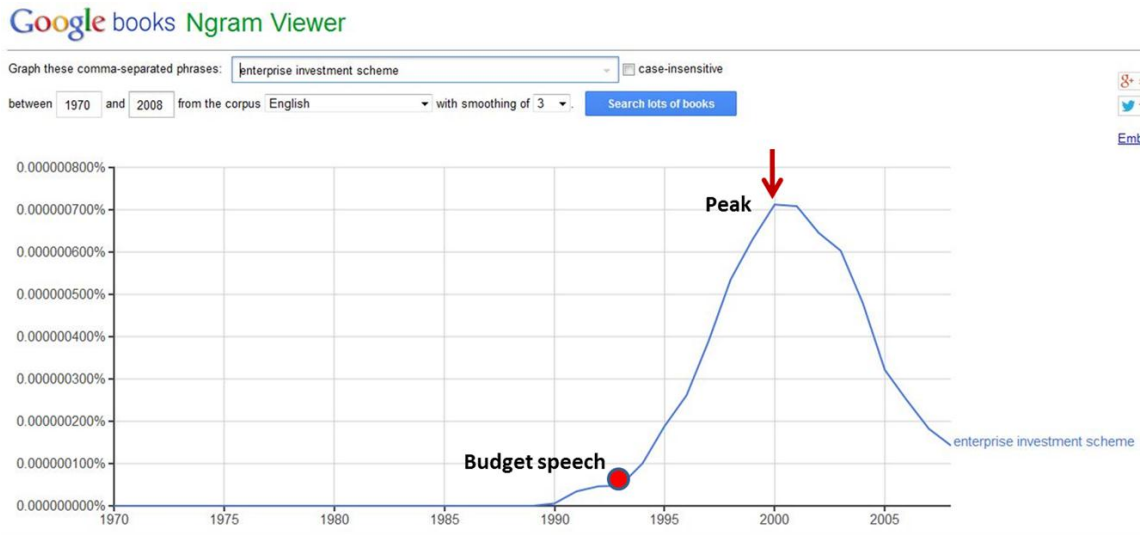

Figure 9. Budget_1994_Clarke
3-gram keyword: peak in 2000, after 6 ys from speech

6. Conclusion

In this work, we tried to propose a mixed model analysis for extracting keywords in a specific domain, in this case UK budget speeches. The model is applicable to any field of speech and would help us make generalization about the emergence of keywords; their syntactic structure and the frequency of their occurrence. First phase of analysis was an explanatory cluster analysis which provided us with a bird‟s eye view of the corpus. Considering that the holistic analysis of the corpus could not provide us with the dominant structural classes of terms, we did a correspondence analysis. A GNEPA method was applied for extracting the frequent combinatory classes in the corpus. To get a better view of collocations , also for being able to extract the keywords, we visualized the data using Vos-viewer and Pajek. The graph-based mapping of significant collocations which also depicts betweenness, closeness and degree centralities better shows the position and centrality of keywords in a term cluster. Graph-based mapping of significant collocations not only allows for a more precise statistical analysis of the data, but also increases the explanatory power of the model.

References

Das, B., Pal, S., Mondal, S. K., Dalui, D., & Shome, S. K. (2013). Automatic keyword extraction from any text document using N-gram rigid collocation. Int. J. Soft Comput. Eng.(IJSCE), 3(2), 238-242.

Firth, J. R. (1957). Papers in linguistics, 1934-1951. London; New York, Oxford University Press.

Hausmann, F. J. (1989). Le dictionnaire de collocations. *FJ Hausmann et al*, 1010-1019.

Leydesdorff, L. (2009). "How are new citation-based journal indicators adding to the bibliometric toolbox?" *Journal of the American Society for Information Science and Technology* 60(7): 1327-1336.

Markov, A. A., & Liebmann, H. (1912). Wahrscheinlichkeitsrechnung. Leipzig, Berlin: B.G. Teubner.

Michel* Jean-Baptiste et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331 (2011).

Nesselhauf, N. (2005). *Collocations in a learner corpus* (Vol. 14). John Benjamins Publishing.

Ratinaud, P., & Marchand, P. (2012). Application de la méthode ALCESTE à de" gros" corpus et stabilité des" mondes lexicaux": analyse du" Cable-Gate" avec IraMuTeQ. Actes des 11e Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012.

Rogers, E. M. (1962). *Diffusion of innovations*. New York: Free Press of Glencoe.

Seretan, V. (2011). Syntax-based collocation extraction (Vol. 44). Springer Science & Business Media.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, vol.30: 50–64.

Siepmann, D. (2005). Collocation, colligation and encoding dictionaries. Part I: Lexicological Aspects. International Journal of Lexicography, 18(4), 409-443.

Van Eck et al. (2009). An Experimental Comparison of Bibliometric Mapping Techniques. *10th International Conference on Science and Technology Indicators.* Vienna, September 18, 2008

Van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *Centre for Science and Technology Studies*, Leiden University.

Wray, A. (2002). Formulaic language and the lexicon. Cambridge; New York, Cambridge University Press.